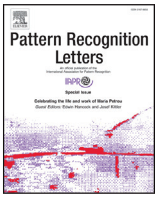
Pattern Recognition Letters 000 (2018) 1–9

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Pattern Recognition Letters

Review on mining data from multiple data sources

Ruili Wang[a], Wanting Ji[a,*], Mingzhe Liu[b], Xun Wang[c], Jian Weng[d], Song Deng[e], Suying Gao[f], Chang-an Yuan[g]

[a] *Institute of Natural and Mathematical Sciences, Massey University, Room 2.14, Mathematical Sciences Building, Auckland 0632, New Zealand*
[b] *School of Network Security, Chengdu University of Technology, Chengdu 610059, China*
[c] *School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*
[d] *College of Cyber Security, Jian University, Guangzhou 519632, China*
[e] *School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210043, China*
[f] *School of Economics and Management, Hebei University of Technology, Tianjin 300401, China*
[g] *School of Computer and Information Engineering, Guangxi Teachers Education University, Nanning 530023, China*

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Multiple data source mining
Pattern analysis
Data classification
Data clustering
Data fusion

ABSTRACT

In this paper, we review recent progresses in the area of mining data from multiple data sources. The advancement of information communication technology has generated a large amount of data from different sources, which may be stored in different geological locations. Mining data from multiple data sources to extract useful information is considered to be a very challenging task in the field of data mining, especially in the current big data era. The methods of mining multiple data sources can be divided mainly into four groups: (i) pattern analysis, (ii) multiple data source classification, (iii) multiple data source clustering, and (iv) multiple data source fusion. The main purpose of this review is to systematically explore the ideas behind current multiple data source mining methods and to consolidate recent research results in this field.

1. Introduction

The advancement of information communication technology has generated a large amount of data from different sources, which may be stored in different geological locations. Each database may have its own structure to store data. Mining multiple data sources [1–3] distributed at different geological locations to discover useful patterns are critical important for decision making. In particular, the Internet can be seen as a large, distributed data repository consisting of a variety of data sources and formats, which can provide abundant information and knowledge.

Data from different sources may seem irrelevant to each other. Once information generated from different sources is integrated, new and useful knowledge may emerge. Here is an excellent example of how an organization to utilize mining data from different data sources to obtain profound information, which cannot obtain from an individual source.

The Australian Taxation Office (ATO) mines data from different data sources such as social media posts, private school records and immigration data to detect tax cheats. Mining data from different data sources become a sophisticated tool to crackdown tax cheats that yielded nearly $10 billion in 2016 [4]. For example, in a normal Australian family, the husband has a business and reported $80,000 of taxable income per year, putting him just inside the second-lowest tax bracket, and his wife reported earning $60,000 per year. But the data collected from different data sources revealed that the family had three children at private schools at an estimated cost of $75,000 per year, while immigration records and social media posts showed that the family had recently taken five business-class flights and a holiday in a Canadian ski resort, Whistler. It means their declared incomes did not match their lifestyle. This prompted ATO to contact them to confirm if they have unpaid taxes. From the above example, we can see that developing an effective data mining technique for mining from multiple data sources to discover useful information is crucially important for decision making.

However, how to efficiently mine quality information from multiple data sources is a challenging task for current research [5–9], especially in the current big data era, because in real world applications, data stored in multiple places often have conflictions [10]. The conflictions include: (i) data name conflictions: (a) the same object has different names in different data sources, or (b) two different objects from different data sources may have the same name; (ii) data format conflictions: the same object in different data sources has different data types, domains, scales, and preci-

[*] Corresponding author.
E-mail address: jwt@escience.cn (W. Ji).

https://doi.org/10.1016/j.patrec.2018.01.013
0167-8655/© 2018 Elsevier B.V. All rights reserved.

sions; (iii) data value confliction: the same object in different data sources records different values; (iv) data sources confliction: different data sources have different database structures.

In order to overcome these conflictions, four effective approaches have been adopted to mine useful information and discover new knowledge from multiple data sources: (i) pattern analysis [11–14], which mining useful patterns and information from one data source or several data sources in accordance with changing conditions constraints or relationships; (ii) multiple data source classification, which labels data sources according to a certain standard, then classifies them by their labels; (iii) multiple data source clustering, which clusters data sources according to their similarities/distances; (iv) multiple data source fusion, which combines data from multiple data sources to achieve higher accuracy and more specific ratiocinations. Based on these three approaches, we can mine useful information from multiple data sources to discover new knowledge according to individual needs.

The main purpose of this review is to systematically overview current multiple source data mining methods and to consolidate recent research results in this field. In this paper, the pattern analysis approaches for multiple data sources are reviewed first, and the approach of multiple data source classification and clustering is reviewed after that. Then, mining multiple data sources using data fusion is reviewed. The rest of this paper is organized as follows. Section 2 shows typical methods of multiple data source mining using pattern analysis, while Section 3 is about multiple data source classification and clustering. Section 4 describes methods of multiple data source fusion. Finally, Section 5 provides conclusions and discussions.

2. Key methods for pattern analysis

In this section, we will describe different methods of pattern analysis on multiple data sources. Pattern analysis on multiple data sources is a process of mining valuable information or extracting hidden predictive information from different databases. Patterns existing in multiple data sources can be categorized as local patterns and global patterns. A pattern identified from a mono-database can be seen as a local pattern, while a global pattern has to be determined based on all the obtained local patterns from multiple data sources [15–17]. Correspondingly, two mainstream methodologies of pattern analysis are developed, named local pattern analysis and global pattern analysis, which will be discussed as follows.

2.1. Local pattern analysis

Local patterns analysis is a process of identifying patterns from a mono-database, which can efficiently extract knowledge from the mono-database [18]. Based on previous studies [19–25], we can summarize the local pattern analysis methods into three categories: association rule mining, sequential pattern mining, and others.

2.1.1. Association rule mining

Association rule mining is a process of discovering the probability of the co-occurrence of items in a database. Association rules express the relationships between co-occurring items, which are often used to analyze sales transactions. Association rule mining is also known as market basket analysis. There are four parameters used to describe an association rule: support, confidence, expected confidence, and lift.

For example, a transaction database of a supermarket records the number of customers on a given day is 1000. Among them, there are 100 customers bought both bread and milk, the support of bread → milk (bread and milk are two itemsets, bread → milk presents an association rule between bread and milk) is 10%. If 40% of customers who bought bread will also buy milk, then the confidence of bread → milk is 40%. If there are totally 200 customers bought milk in the given day, then the expected confidence of bread → milk is 20%. Thus, the lift of bread → milk is $40\%/20\% = 2$.

While the support measures the importance of an association rule, the confidence measures the accuracy of an association rule. The support can reflect the representativeness of an association rule. Although some association rules have high confidences, their supports are very low, which means these association rules have little chance of being practical so that they are also unimportant.

The expected confidence of an association rule can be seen as the support of an itemset without the influence of other itemsets, while the lift of an association rule is the ratio of confidence to expected confidence, which is used to describe how one itemset affects another itemset. In general, the lift of a useful association rule should be greater than 1. Only when the confidence of an association rule is greater than its expected confidence, it shows a correlation between these two itemsets (i.e., one itemset has a promoting effect on another itemset), otherwise, this association rule is meaningless.

According to the definition above, any two itemsets in a transaction can have an association rule between them. In fact, users are only interested in the association rules that satisfy a certain support and confidence. Therefore, a support-confidence framework is used to mine the correlations between itemsets in the local databases. Two thresholds, the minimum support and the minimum confidence, specify the minimum requirements of a meaningful association rule.

Association rule mining consists of two phases: (i) In the first phase, all the high frequent itemsets are identified. (ii) In the second phase, association rules are generated from the high frequent itemsets. It has been extensively studied to identify the relation existing in databases between itemsets [26].

A classical algorithm of association rule mining named Apriori algorithm is presented by Agrawal and Srikant [27], which is used for mining frequent itemsets and relevant association rules. A key concept in the Apriori algorithm is the anti-monotonicity of the support measure. It assumes that: (i) All subsets of a frequent itemset must be frequent. (ii) For any infrequent itemset, all its supersets must be infrequent. The algorithm can be divided into two phases. In the first phase, it applies minimum support to find all the frequent sets with k items in a database. In the second phase, it uses the self-join rule to find the frequent sets of $k+1$ items with the help of frequent k-itemsets. Repeat this process from $k=1$ to the point when we are unable to apply the self-join rule. The Apriori algorithm can find the relationship between items in a database and it is the first efficient and practical algorithm for association rule mining. But it needs to generate a large number of candidate itemsets and repeatedly scan the entire database.

A Frequent-Pattern tree structure called FP-tree is presented by Han et al. in [28], which is constructed by (i) Scan a database to get all the frequent itemsets and their respective support in the database, sorting frequent items in frequent itemsets in a descending order of their support values. (ii) Create the root node and scan the database again. Then select the frequent items and sort them in the order of frequent itemsets. It is used for collecting compacted and important information about frequent patterns. According to FP-tree, a corresponding method called frequent-pattern growth algorithm is proposed for mining the complete set of frequent patterns [28]. The proposed method can efficiently avoid the process of candidate generation-and-test, and it can cut down the time cost of frequent pattern mining. However, the algorithm may have a challenge when encountering graph-based patterns or noise samples.

Please cite this article as: R. Wang et al., Review on mining data from multiple data sources, Pattern Recognition Letters (2018), https://doi.org/10.1016/j.patrec.2018.01.013